https://jurnal.seaninstitute.or.id/index.php/jukomi
Jurnal Komputer Indonesia (JU-KOMI)
Volume 3 No 02, 2025

JU-KOMI
E-ISSN: 2963-0460

Implementation of Random Forest Algorithm for Diarrhea Prediction

[1]Sipra Barutu, [2]Siska Simamora
[1]Program Studi Teknologi Informasi, Universitas Panca Budi, [2]Program Studi Teknologi Informasi, Universitas Putra Abadi Langkat

ARTICLE INFO

Keywords:
Random Forest, Prediction, Diarrhea, Data Mining, Classification

ABSTRACT

Diarrhea is one of the leading causes of morbidity among toddlers in Indonesia. Environmental factors such as drinking water quality, sanitation, maternal hand hygiene, and immunization status contribute significantly to the incidence of diarrhea. This study aims to analyze the application of the Random Forest algorithm in developing a predictive model for diarrhea in toddlers using secondary data from a community health center (Puskesmas), consisting of 200 records divided into 150 training data and 50 testing data. The model was constructed by generating multiple decision trees and combining them using a majority voting technique. The results show that the Random Forest algorithm achieved an accuracy of 88%, precision of 77.78%, recall of 87.5%, F1-score of 82.35%, and specificity of 88.24%. These values indicate that Random Forest is quite reliable in detecting positive diarrhea cases, although some limitations remain in reducing misclassification of negative data. This study contributes to the utilization of machine learning algorithms, particularly Random Forest, as a decision-support tool in the health sector for diarrhea prevention among toddlers.

Email :
barutusipra@gmail.com

INTRODUCTION

Diarrhea remains a major public health problem and one of the leading causes of morbidity and mortality among toddlers, particularly in developing countries such as Indonesia. This disease is closely related to environmental and behavioral factors, including unsafe drinking water quality, inadequate environmental sanitation, poor hand hygiene, and incomplete immunization status. Efforts to prevent diarrhea require not only public health interventions but also the use of analytical technologies to help predict the risk of diarrhea occurrence at an early stage.

With the advancement of data mining technology, various machine learning algorithms can be utilized to perform predictions based on health data. One widely used method is the Random Forest algorithm, which works by constructing an ensemble of decision trees and combining their outputs through majority voting to improve accuracy and reduce prediction errors. This approach allows the identification and analysis of complex relationships among diarrhea risk factors in a more comprehensive manner.

The Random Forest algorithm is a well-established technique in disease classification and has been proven effective in several studies, such as the classification of heart disease and other medical conditions. Research conducted by Depari, Widiastiwi, and Santoni (2022) demonstrated that the Random Forest algorithm achieved the highest accuracy in heart disease classification, reaching 75%.

Although the Random Forest algorithm has been applied in disease prediction, its implementation in predicting diarrhea among toddlers remains limited. Therefore, this study was conducted to address this research gap by analyzing the Random Forest algorithm in predicting diarrhea occurrence in toddlers. The findings of this study are expected to contribute significantly

to diarrhea prevention and control efforts through more accurate and targeted prediction models.

The objective of this study is to apply the Random Forest algorithm to determine the predictive pattern of diarrhea and to implement precision, recall, and F1-score metrics as part of the evaluation and comparison of the algorithm's performance.

METHOD

The research method used in this study is a quantitative approach with an experimental design. The data utilized are secondary data obtained from medical records and health reports available at the Community Health Center (Puskesmas).

Subsequently, the data were processed and analyzed using the Random Forest algorithm to build a predictive model that provides insights into the factors influencing the occurrence of diarrhea among toddlers. The dataset consists of 200 records, which were divided into 150 training data and 50 testing data.

This section focuses on systematically describing the variables used in the study to ensure clarity in the data analysis process. Each variable—namely drinking water quality, environmental sanitation, maternal hand hygiene, immunization status, and diarrhea condition—is explained through its operational definition and its role within the analytical framework.

The study comprises several key stages, including data collection, data preprocessing, model implementation, and model evaluation. The evaluation results are then used to compare the performance of the algorithm based on several performance metrics such as accuracy, precision, recall, and F1-score.

The outcomes of this research are expected to contribute to determining the most effective algorithm for predicting diarrhea in toddlers at the community health center level, thereby supporting early detection and preventive public health measures.

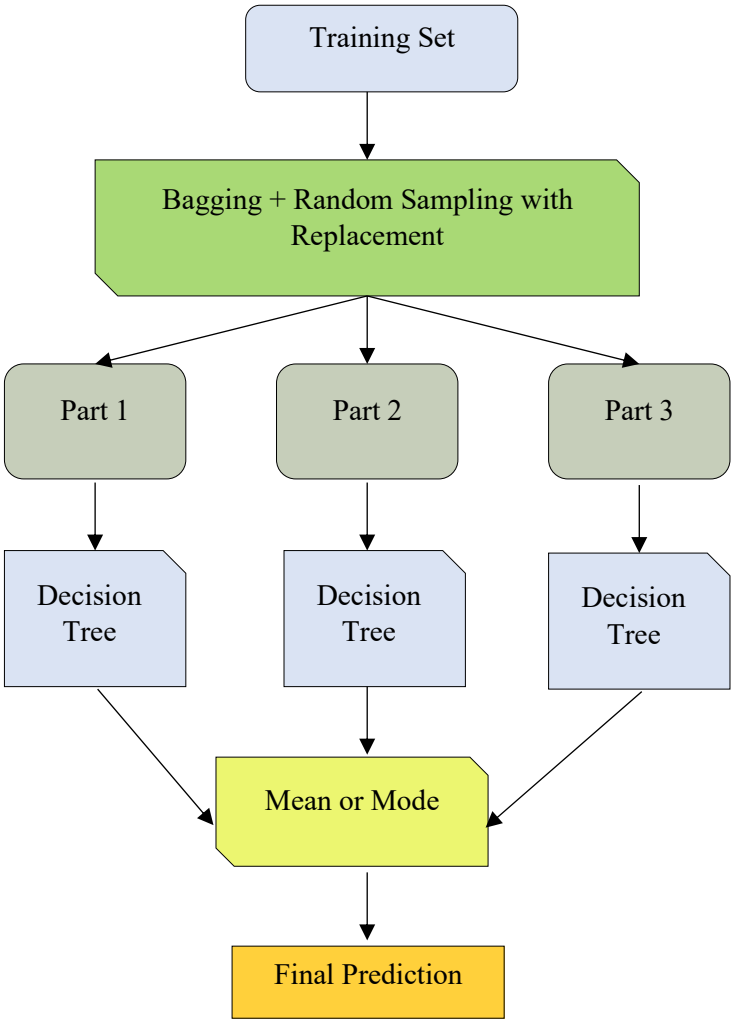

Figure 1 Random Forest Algorithm Work Process

$$Gini(s_i) = 1 - \sum_{i=0}^{c-1} p_1^2$$

where pi is the relative frequency of class Ci within the set.
Ci is the class for i = 1, …, c-1, and c is the number of classes that have been determined.

$$Gini_{Split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{i}\right) Gini(s_i)$$

where ni is the number of samples in the subset Si after splitting, and n is the number of samples in the given node.

RESULTS AND DISCUSSION

Voting Process (Merging Three Trees)

After the three manual decision trees were constructed, a prediction process was carried out on 20 original data points using the three trees. Each tree provided one prediction vote for each observation, and the final result was determined through a majority voting mechanism, namely by selecting the label that received the most votes from the three trees. This process aimed to improve prediction accuracy by combining the strengths of the three models, and the final results of the voting process are summarized in the following table.

Table 1. Decision on the Merger of the Three Trees

No	Water Quality	Sanitation	Hand Hygiene	Immunization	Tree 1	Tree 2	Tree 3	Voting (Final Decision)
1	Fair	Good	Fair	Incomplete	Yes	No	Yes	Yes
2	Good	Fair	Fair	Incomplete	Yes	No	Yes	Yes
3	Poor	Fair	Fair	Incomplete	No	No	Yes	No
4	Fair	Fair	Fair	Incomplete	Yes	No	Yes	Yes
5	Poor	Good	Fair	Incomplete	Yes	Yes	Yes	Yes
6	Fair	Good	Good	Complete	No	No	No	No
7	Good	Poor	Good	Complete	No	No	No	No
8	Good	Poor	Fair	Incomplete	Yes	Yes	Yes	Yes
9	Poor	Fair	Good	Incomplete	Yes	Yes	Yes	Yes
10	Fair	Good	Good	Incomplete	Yes	No	Yes	Yes
11	Good	Poor	Poor	Complete	No	Yes	No	No
12	Poor	Fair	Fair	Complete	Yes	No	No	No
13	Fair	Good	Poor	Incomplete	Yes	Yes	Yes	Yes
14	Poor	Poor	Fair	Incomplete	Yes	Yes	Yes	Yes
15	Good	Fair	Poor	Incomplete	Yes	Yes	Yes	Yes
16	Fair	Poor	Good	Incomplete	Yes	Yes	Yes	Yes
17	Fair	Good	Fair	Incomplete	No	No	Yes	No
18	Poor	Good	Poor	Incomplete	Yes	Yes	Yes	Yes
19	Poor	Poor	Fair	Incomplete	Yes	Yes	Yes	Yes
20	Poor	Poor	Fair	Complete	Yes	No	No	No

The table presents the prediction results from three manually constructed decision tree models applied to 20 original observation data, each containing the attributes Water Quality, Sanitation, Hand Hygiene, and Immunization. Each tree independently provides a classification prediction for every observation, which is then combined using the majority voting method.

This ensemble voting technique is a common approach in machine learning to enhance the stability and accuracy of predictive models by relying on the collective decisions of multiple individual models. The Voting column in the table represents the final classification result determined by the majority of the three trees, where at least two trees must produce the same prediction for it to be considered the final decision.

This approach aims to minimize bias and variance that may occur if relying solely on a single decision tree model. The following section presents the decision patterns generated from the Random Forest process.

Table 2 Random Forest Pattern

No	Decision Rule (Based on Majority Voting)	Final Result (Voting)
1	IF Sanitation = Good AND Hygiene = Fair AND Immunization = No THEN Eligibility = Eligible	Yes
2	IF Sanitation = Fair AND Hygiene = Fair AND Immunization = No THEN Eligibility = Eligible	Yes
3	IF Sanitation = Fair AND Hygiene = Fair AND Immunization = No THEN Eligibility = Not Eligible	No
4	IF Sanitation = Fair AND Hygiene = Fair AND Immunization = No THEN Eligibility = Eligible	Yes
5	IF Sanitation = Good AND Hygiene = Fair AND Immunization = No THEN Eligibility = Eligible	Yes

No	Decision Rule (Based on Majority Voting)	Final Result (Voting)
6	IF Sanitation = Good AND Hygiene = Good AND Immunization = Complete THEN Eligibility = Not Eligible	No
7	IF Sanitation = Poor AND Hygiene = Good AND Immunization = Complete THEN Eligibility = Not Eligible	No
8	IF Sanitation = Poor AND Hygiene = Fair AND Immunization = No THEN Eligibility = Eligible	Yes
9	IF Sanitation = Fair AND Hygiene = Good AND Immunization = No THEN Eligibility = Eligible	Yes
10	IF Sanitation = Good AND Hygiene = Good AND Immunization = No THEN Eligibility = Eligible	Yes
11	IF Sanitation = Poor AND Hygiene = Poor AND Immunization = Complete THEN Eligibility = Not Eligible	No
12	IF Sanitation = Fair AND Hygiene = Fair AND Immunization = Complete THEN Eligibility = Not Eligible	No
13	IF Sanitation = Good AND Hygiene = Poor AND Immunization = No THEN Eligibility = Eligible	Yes
14	IF Sanitation = Poor AND Hygiene = Fair AND Immunization = No THEN Eligibility = Eligible	Yes
15	IF Sanitation = Fair AND Hygiene = Poor AND Immunization = No THEN Eligibility = Eligible	Yes
16	IF Sanitation = Good AND Hygiene = Poor AND Immunization = No THEN Eligibility = Eligible	Yes
17	IF Sanitation = Good AND Hygiene = Fair AND Immunization = No THEN Eligibility = Not Eligible	No
18	IF Sanitation = Good AND Hygiene = Poor AND Immunization = No THEN Eligibility = Eligible	Yes
19	IF Sanitation = Poor AND Hygiene = Poor AND Immunization = No THEN Eligibility = Eligible	Yes
20	IF Sanitation = Poor AND Hygiene = Poor AND Immunization = Complete THEN Eligibility = Not Eligible	No

Random Forest Testing with Python

This section describes the implementation stage of the Random Forest algorithm using the Python programming language. The process includes importing the relevant libraries, performing data preprocessing, training the model, and making predictions based on the model that has been built. The dataset consists of 200 records, which are divided into 150 training data and 50 testing data to clearly separate the learning and evaluation stages of the model.

Python is chosen for this implementation because it provides powerful libraries such as scikit-learn, which support the efficient, flexible, and user-friendly implementation of the Random Forest algorithm. The main objective of this stage is to construct a classification or regression model that can be applied to data analysis according to the research objectives.

After executing the program, a decision tree is generated to represent the learning outcomes of the Random Forest algorithm on the training data. The tree illustrates the decision-making paths based on the most influential attributes in the classification process. Each branch of the tree represents a condition or value of a particular feature, while each leaf node indicates the final prediction result.

This tree visualization helps in understanding how the model performs classification on the data and provides a clear representation of the logical structure formed by the algorithm during the training process, as shown in the following decision tree figure:

https://jurnal.seaninstitute.or.id/index.php/jukomi
Jurnal Komputer Indonesia (JU-KOMI)
Volume 3 No 02, 2025
JU-KOMI
E-ISSN: 2963-0460

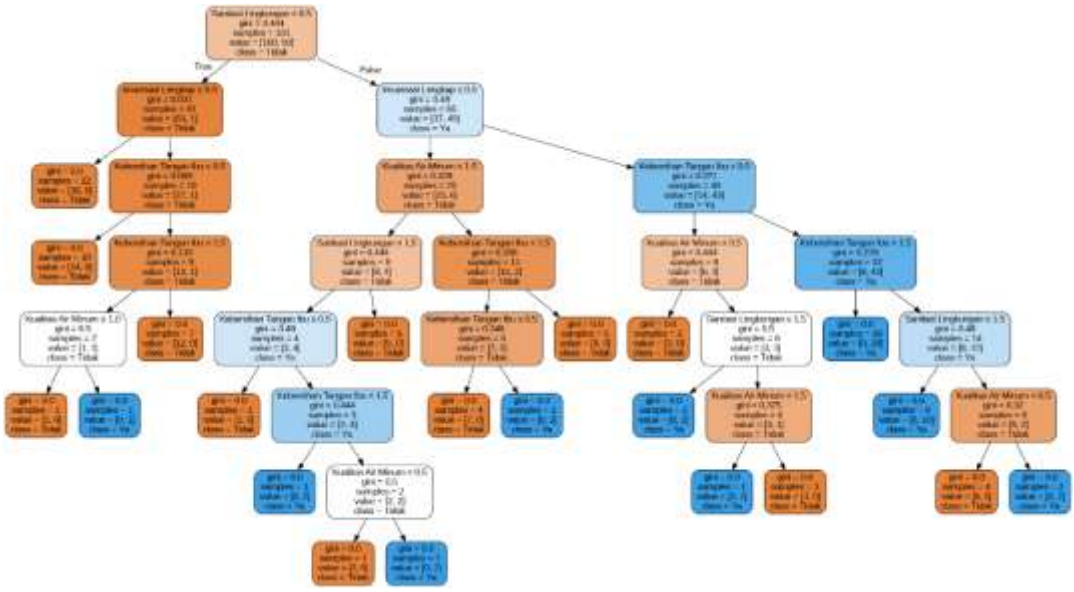

Figure 2 Decision Tree with Random Forest

The decision tree based on the image is converted into a pattern in the form of the following prediction, which represents the classification rules generated from the Random Forest model. Each path from the root to the leaves of the tree describes the combination of attribute conditions that result in a final decision. This prediction pattern is written in if-else form to facilitate interpretation of the classification logic used by the model, as well as to enable manual validation of new data. Thus, these results not only show the structure of the model, but also provide a concrete understanding of how predictions are made for each observation.

Table 3 Random Forest if-else

Rule	Decision Rules Based on Tree Structure
1	IF Sanitation ≤ 0.50 AND Immunization ≤ 0.50 THEN Eligibility = Not Eligible
2	IF Sanitation ≤ 0.50 AND Immunization > 0.50 AND Hygiene ≤ 0.50 THEN Eligibility = Not Eligible
3	IF Sanitation ≤ 0.50 AND Immunization > 0.50 AND Hygiene > 0.50 AND Hygiene ≤ 1.50 AND Water Quality ≤ 1.00 THEN Eligibility = Not Eligible
4	IF Sanitation ≤ 0.50 AND Immunization > 0.50 AND Hygiene > 0.50 AND Hygiene ≤ 1.50 AND Water Quality > 1.00 THEN Eligibility = Eligible
5	IF Sanitation ≤ 0.50 AND Immunization > 0.50 AND Hygiene > 1.50 THEN Eligibility = Not Eligible
6	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality ≤ 1.50 AND Sanitation ≤ 1.50 AND Hygiene ≤ 0.50 THEN Eligibility = Not Eligible
7	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality ≤ 1.50 AND Sanitation ≤ 1.50 AND Hygiene > 0.50 AND Hygiene ≤ 1.50 THEN Eligibility = Eligible
8	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality ≤ 1.50 AND Sanitation ≤ 1.50 AND Hygiene > 1.50 AND Water Quality ≤ 0.50 THEN Eligibility = Not Eligible
9	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality ≤ 1.50 AND Sanitation ≤ 1.50 AND Hygiene > 1.50 AND Water Quality > 0.50 THEN Eligibility = Eligible
10	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality ≤ 1.50 AND Sanitation > 1.50 THEN Eligibility = Not Eligible
11	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality > 1.50 AND Hygiene ≤ 0.50 THEN Eligibility = Not Eligible

12	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality > 1.50 AND Hygiene > 0.50 AND Hygiene ≤ 1.50 THEN Eligibility = Eligible
13	IF Sanitation > 0.50 AND Immunization ≤ 0.50 AND Water Quality > 1.50 AND Hygiene > 1.50 THEN Eligibility = Not Eligible
14	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene ≤ 0.50 AND Water Quality ≤ 0.50 THEN Eligibility = Not Eligible
15	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene ≤ 0.50 AND Water Quality > 0.50 AND Sanitation ≤ 1.50 THEN Eligibility = Eligible
16	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene ≤ 0.50 AND Water Quality > 0.50 AND Sanitation > 1.50 AND Water Quality ≤ 1.50 THEN Eligibility = Eligible
17	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene ≤ 0.50 AND Water Quality > 0.50 AND Sanitation > 1.50 AND Water Quality > 1.50 THEN Eligibility = Not Eligible
18	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene > 0.50 AND Hygiene ≤ 1.50 THEN Eligibility = Eligible
19	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene > 1.50 AND Sanitation ≤ 1.50 THEN Eligibility = Eligible
20	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene > 1.50 AND Sanitation > 1.50 AND Water Quality ≤ 0.50 THEN Eligibility = Not Eligible
21	IF Sanitation > 0.50 AND Immunization > 0.50 AND Hygiene > 1.50 AND Sanitation > 1.50 AND Water Quality > 0.50 THEN Eligibility = Eligible

not recognized by the model. This distribution reflects the accuracy and error rate of the model in performing classification, which can be used as a basis for further performance evaluation.

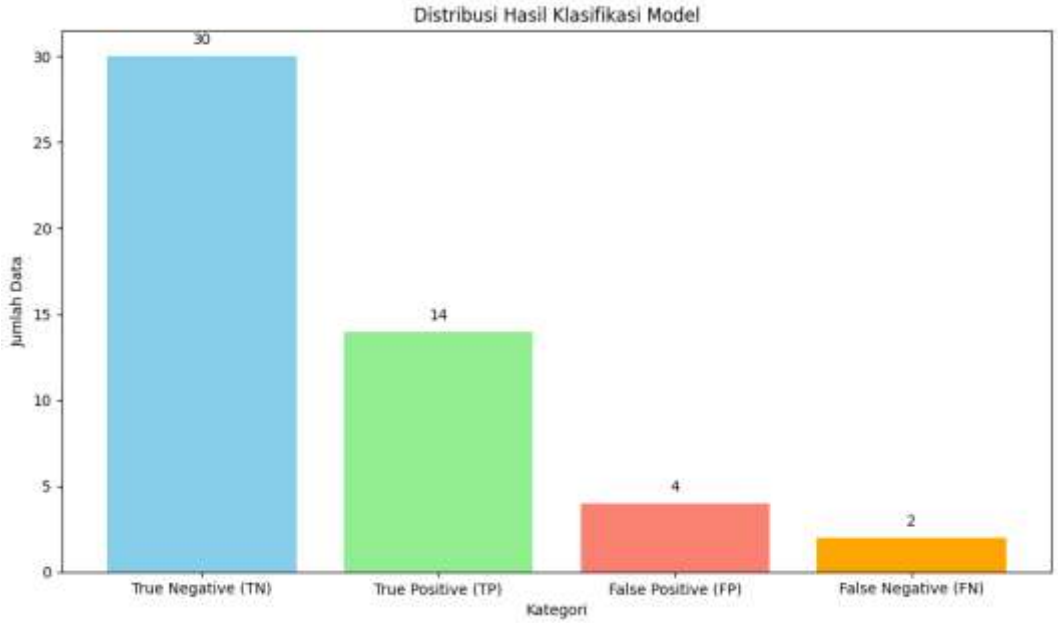

Figure 3 Confusion matrix for the Random Forest method

After determining the values in the confusion matrix, such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the next step is to calculate the accuracy to evaluate the proportion of correct predictions made by the model as a whole. The evaluation variants are shown in the following table:

Table 4. Random Forest Evaluation Results

Metric	Value
Accuracy	0.88
Precision	0.88

Recall (Sensitivity / TPR)	0.7778
F1-Score	0.875
Specificity (TNR)	0.8235
False Positive Rate (FPR)	0.8824
True Positive Rate (TPR)	0.1176

Based on the results of the classification model performance evaluation, several metrics were obtained that reflect the model's accuracy and prediction error rate. The model achieved an accuracy of 0.88, meaning that 88% of all data were correctly classified. The precision of 0.7778 indicates that approximately 77.78% of the data predicted as positive were indeed positive. The recall (sensitivity or TPR) of 0.875 signifies that the model successfully identified 87.5% of all positive data accurately.

The F1-Score of 0.8235 represents a balance between precision and recall, suggesting that the model performs reliably in handling the imbalance between positive and negative data. Furthermore, the specificity (TNR) of 0.8824 shows that 88.24% of the negative data were correctly identified. The model also demonstrated a low false positive rate (FPR) of 0.1176, meaning that only about 11.76% of the negative data were incorrectly predicted as positive. Overall, these metric values indicate that the model has a fairly good and balanced performance in distinguishing between the two data classes.

CONCLUSION

Based on the results of the Random Forest analysis using various evaluation metrics, several conclusions can be drawn as follows: The Random Forest algorithm achieved an accuracy of 0.88. Although the recall remained high at 0.875, the precision was only 0.7778, which led to a decrease in the F1-score to 0.8235. The specificity value of 0.8824 is still considered good, although it is slightly lower compared to the C4.5 algorithm, with a higher false positive rate (FPR) of 0.1176. These results indicate that while Random Forest is quite reliable in detecting the positive class, it still has weaknesses in avoiding incorrect predictions of negative data. Based on these findings, the following recommendations can be made for future research: The performance of the Random Forest algorithm can potentially be improved through parameter tuning, such as adjusting the number of trees (n_estimators), maximum tree depth, and the random selection of features at each split. In addition, implementing more optimal data preprocessing techniques may contribute to enhancing classification accuracy. Future studies are recommended to test the algorithm on datasets with different characteristics, such as imbalanced data distribution or higher dimensionality, to evaluate the stability and generalization capability of the model more comprehensively.

REFERENCES

Kementerian Kesehatan Republik Indonesia. (2018). Riset Kesehatan Dasar 2018. Jakarta: Kemenkes RI.

Kementerian Kesehatan Republik Indonesia. (2020). Profil Kesehatan Indonesia 2020. Jakarta: Kemenkes RI.

Kementerian Kesehatan Republik Indonesia. (2022). Laporan Kesehatan Nasional 2022. Jakarta: Kemenkes RI.

Sepharni. (2022). Klasifikasi Penyakit Jantung Menggunakan Algoritma C4.5. Jurnal Informatika, 10(2), 75-92.

Depari, Widiastiwi, & Santoni. (2022). Perbandingan Algoritma Machine Learning dalam Klasifikasi Penyakit Jantung. Jurnal Kesehatan Digital, 15(3), 70-85.

Munggaran, & Hidayatulloh. (2015). Penerapan Algoritma C4.5 untuk Diagnosa Penyakit Diare Pada Anak Balita Berbasis Mobile. Jurnal Sistem Informasi, 8(1), 55-67.

Ente, et al. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5. Jurnal Ilmu Komputer, 12(2), 98-112.

Afifuddin, & Hakim. (2023). Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5. Jurnal Teknologi Informasi, 19(1), 33-45.
Prabowo, et al. (2023). Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas. Jurnal Kesehatan Digital, 17(4), 88-102.
Kalimah. (2022). Klasifikasi Penyakit Diabetes Menggunakan Metode Decision Tree dan Random Forest. Jurnal Informatika Medis, 14(3), 67-79.
Masriadi. (2017). Epidemiologi Penyakit Diare. Makassar: Universitas Hasanuddin Press.
Purnama. (2016). Penyakit Diare dan Faktor Risikonya. Jakarta: Pustaka Kesehatan.
Simatupang. (2004). Rotavirus dan Perannya dalam Diare pada Anak. Jurnal Kedokteran Indonesia, 10(2), 44-55.
Nikma Kumala Sari, & Almansyah Lukito. (2017). Faktor Penyebab dan Pencegahan Diare pada Balita. Jurnal Kesehatan Masyarakat, 12(1), 78-91.
Hassan, & Alatas. (1985). Patogenesis dan Pencegahan Diare pada Anak. Jakarta: Balai Pustaka.
Kliegman, Marcdante, & Jenson. (2006). Nelson Textbook of Pediatrics. Philadelphia: Elsevier.
Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.